Unsupervised Tracklet Person Re-Identification

Minxian Li, Xiatian Zhu, and Shaogang Gong

Abstract—Most existing person re-identification (re-id) methods rely on *supervised* model learning on per-camera-pair *manually* labelled pairwise training data. This leads to poor scalability in a practical re-id deployment, due to the lack of exhaustive identity labelling of positive and negative image pairs for every camera-pair. In this work, we present an unsupervised re-id deep learning approach. It is capable of incrementally discovering and exploiting the underlying re-id discriminative information from *automatically* generated person tracklet data *end-to-end*. We formulate an *Unsupervised Tracklet Association Learning* (UTAL) framework. This is by jointly learning within-camera tracklet discrimination and cross-camera tracklet association in order to maximise the discovery of tracklet identity matching both within and across camera views. Extensive experiments demonstrate the superiority of the proposed model over the state-of-the-art unsupervised learning and domain adaptation person re-id methods on eight benchmarking datasets.

Index Terms—Person Re-Identification; Unsupervised Tracklet Association; Trajectory Fragmentation; Multi-Task Deep Learning.

1 Introduction

Person re-identification (re-id) aims to match the underlying identity classes of person bounding box images detected from non-overlapping camera views [1]. In recent years, extensive research has been carried out on re-id [2,3,4,5,6,7]. Most existing person re-id methods, in particular neural network deep learning models, adopt the *supervised learning* approach. Supervised deep models assume the availability of a large number of *manually* labelled *cross-view identity matching image pairs* for each camera pair. This enables deriving a feature representation and/or a distance metric function optimised for each camera-pair. Such an assumption is inherently limited for generalising a person re-id model to many different camera networks. This is because, exhaustive manual identity (ID) labelling of positive and negative person image pairs for every camera-pair is prohibitively expensive, given that there are a quadratic number of camera pairs in a surveillance network.

It is no surprise that person re-id by *unsupervised learning* become a focus in recent research. In this setting, per-camera-pair ID labelled training data is no longer required [8,9,10,11,12,13,14,15,16]. However, existing unsupervised learning re-id models are significantly inferior in re-id accuracy. This is because, lacking cross-view pairwise ID labelled data deprives a model's ability to learn strong discriminative information. This nevertheless is critical for handling significant appearance change across cameras.

An alternative approach is to leverage jointly (1) unlabelled data from a target domain which is freely available, e.g. videos of thousands of people travelling through a camera view everyday in a public scene, and (2) pairwise ID labelled datasets from independent source domains [17,18,19,20]. The main idea is to first learn a "view-invariant" representation from ID labelled source data, then adapt the pre-learned model to a target domain by using only unlabelled target data. This approach makes an implicit assumption that, the source and target domains share some common cross-view characteristics so that a view-invariant representation can be estimated. This is not always true.

In this work, we consider a *pure* unsupervised person re-id deep learning problem. That is, no ID labelled training data are assumed, neither cross-view nor within-view ID labelling. Although this learning objective shares some modelling spirit with two recent domain transfer models [17,19], both those models do require *suitable* person ID labelled source domain training data, i.e. visually similar to the target domain. Specifically, we consider unsupervised re-id model learning by jointly optimising unlabelled person tracklet data *within-camera* view to be more discriminative and *cross-camera* view to be more associative *end-to-end*.

Our **contributions** are: We formulate a novel unsupervised person re-id deep learning method using automatically generated person tracklets. This avoids the need for camera pairwise ID labelled training data, i.e. *unsupervised tracklet re-id discriminative learning*. Specifically, we propose a ***Unsupervised Tracklet Association Learning*** (UTAL) model with two key ideas: **(1)** *Per-Camera Tracklet Discrimination Learning* that optimises "local" within-camera tracklet label discrimination. It aims to facilitate cross-camera tracklet association given per-camera independently created tracklet label spaces. **(2)** *Cross-Camera Tracklet Association Learning* that optimises "global" cross-camera tracklet matching. It aims to find cross-view tracklet groupings that are most likely of the same person identities without ID label information. This is formulated as to jointly discriminate within-camera tracklet identity semantics and self-discover cross-camera tracklet pairwise matching in end-to-end deep learning. Critically, the proposed UTAL method does not assume any domain-specific knowledge such as camera space-time topology and cross-camera ID overlap. Therefore, it is scalable to arbitrary surveillance camera networks with unknown viewing conditions and background clutters.

Extensive comparative experiments are conducted on

- *Minxian Li is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK. E-mail: m.li@qmul.ac.uk.*
- *Xiatian Zhu is with Vision Semantics Limited, London E1 4NS, UK. E-mail: eddy@visionsemantics.com.*
- *Shaogang Gong is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK. E-mail: s.gong@qmul.ac.uk.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2903058, IEEE Transactions on Pattern Analysis and Machine Intelligence

seven existing benchmarking datasets (CUHK03 [21], Market-1501 [22], DukeMTMC-ReID [23,24], MSMT17 [4], iLIDS-VID [25], PRID2011 [26], MARS [27]) and one newly introduced tracklet person re-id dataset called DukeTracklet. The results show the performance advantages and superiority of the proposed UTAL method over the state-of-the-art unsupervised and domain adaptation person re-id models.

A preliminary version of this work was reported in [28]. Compared with the earlier study, there are a few key differences introduced: **(i)** This study presents a more principled and scalable unsupervised tracklet learning method that learns deep neural network re-id models directly from large scale raw tracklet data. The method in [28] requires a separate preprocessing for domain-specific spatio-temporal tracklet sampling for reducing the tracklet ID duplication rate per camera view. This need for pre-sampling not only makes model learning more complex, not-end-to-end therefore suboptimal, but also loses a large number of tracklets with potential rich information useful for more effective model learning. **(ii)** We propose in this study a new concept of soft tracklet labelling, which aims to explore any inherent space-time visual correlation of the same person ID between unlabelled tracklets within each camera view. This is designed to better address the tracklet fragmentation problem through an end-to-end model optimisation mechanism. It improves person tracking within individual camera views, which is lacking in [28]. **(iii)** Unlike the earlier method, the current model self-discovers and exploits *explicit* cross-camera tracklet association in terms of person ID, improving the capability of re-id discriminative unsupervised learning and leading to superior model performances. **(iv)** Besides creating a new tracklet person re-id dataset, we further conduct more comprehensive evaluations and analyses for giving useful and significant insights.

2 Related Work

Person Re-Identification. Most existing person re-id models are built by *supervised* model learning on a separate set of per-camera-pair ID labelled training data [2,3,4,5,6,7,21,29]. While having no class intersection, the training and testing data are often assumed to be drawn from the same camera network (domain). Their scalability is therefore significantly poor for realistic applications when no such large training sets are available for every camera-pair in a test domain. Human-in-the-loop re-id provides a means of reducing the overall amount of training label supervision by exploring the benefits of human-computer interaction [30,31]. But, it is still labour intensive and tedious. Human labellers need to be deployed repeatedly for conducting similar screen profiling operations whenever a new target domain exhibits. It is therefore not scalable either.

Unsupervised model learning is an intuitive solution to avoiding the need of exhaustively collecting a large set of labelled training data per application domain. However, previous hand-crafted features-based unsupervised learning methods offer significantly inferior re-id matching performance [8,9,10,11,12,14,15,16,32,33], when compared to the supervised learning models. A trade-off between re-id model scalability and generalisation performance can be achieved by semi-supervised learning [13,34]. But these models still assume sufficiently large sized cross-view pairwise ID labelled data for model training.

There are attempts on unsupervised learning by domain adaptation [17,18,19,20,35,36,37]. The idea is to exploit the knowledge of labelled data in "related" source domains through model adaptation on the unlabelled target domain data. One straightforward approach is to convert the source ID labelled training data into the target domain by appearance mimicry. This enables to train a model using the domain style transformed source training data via supervised learning [35,36]. Alternative techniques include semantic attribute knowledge transfer [17,18,38], space-time pattern transfer [39], virtual ID synthesis [40], and progressive adaptation [19,20]. While these models perform better than the earlier generation of methods (Tables 2 and 3), they require similar data distributions and viewing conditions between the labelled source domain and the unlabelled target domain. This restricts their scalability to arbitrarily diverse (and unknown) target domains in large scale deployments.

Unlike all existing unsupervised learning re-id methods, the proposed tracklet association method in this work enables unsupervised re-id deep learning from scratch at end-to-end. This is more scalable and general. Because there is no assumption on either the scene characteristic similarity between source and target domains, or the complexity of handling ID label knowledge transfer. Our method directly learns to discover the re-id discriminative knowledge from *unlabelled* tracklet data automatically generated.

Moreover, the proposed method does not assume any overlap of person ID classes across camera views or other domain-specific information. It is therefore scalable to the scenarios without any knowledge about camera space-time topology [39]. Unlike the existing unsupervised learning method relying on extra hand-crafted features, our model learns tracklet based re-id discriminative features from an end-to-end deep learning process. To our best knowledge, this is the *first* attempt at unsupervised tracklet association based person re-id deep learning model without relying on any ID labelled training video or imagery data.

Multi-Task Learning in Neural Networks. Multi-task learning (MTL) is a machine learning strategy that learns several related tasks simultaneously for their mutual benefits [41]. A good MTL survey with focus on neural networks is provided in [42]. Deep CNNs are well suited for performing MTL. As they are inherently designed to learn joint feature representations subject to multiple label objectives concurrently in multi-branch architectures. Joint learning of multiple related tasks has been proven to be effective in solving computer vision problems [43,44].

In contrast to all the existing methods aiming for supervised learning problems, the proposed UTAL method exploits differently the MTL principle to solve an unsupervised learning task. Critically, our method is uniquely designed to explore the potential of MTL in correlating the underlying *group* level semantic relationships between different individual learning tasks[1]. This dramatically differs from existing MTL based methods focusing on mining the shared knowledge among tasks at the sample level.

[1] In the unsupervised tracklet person re-id context, a group corresponds to a set of categorical labels each associated with an individual person tracklet drawn from a specific camera view.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2903058, IEEE Transactions on Pattern Analysis and Machine Intelligence

Critically, it avoids the simultaneous labelling of multi-tasks on each training sample. Sample-wise multi-task labels are not available in the unsupervised tracklet re-id problem.

Besides, unsupervised tracklet labels in each task (camera view) are *noisy*. As they are obtained without manual verification. Hence, the proposed UTAL model is in effect performing *weakly supervised multi-task learning* with noisy per-task labels. This makes our method fundamentally different from existing MTL approaches that are only interested in discovering discriminative cross-task common representations by *strongly supervised learning* of clean and exhaustive sample-wise multi-task labelling.

Unsupervised Deep Learning. Unsupervised learning of visual data is a long standing research problem starting from the auto-encoder models [45] or earlier. Recently, this problem has regained attention in deep learning. One common approach is by incorporating with data clustering [46] that jointly learns deep feature representations and image clusters. Alternative unsupervised learning techniques include formulating generative models [35], devising a loss function that preserves information flowing [47] or discriminates instance classes [48], exploiting the object tracking continuity cue [49] in unlabelled videos, and so forth.

As opposite to all these methods focusing on uni-domain data distributions, our method is designed particularly to learn visual data sampled from different camera domains with unconstrained viewing settings. Conceptually, the proposed idea of soft tracklet labels is related to data clustering. As the per-camera affinity matrix used for soft label inference is related to the underlying data cluster structure. However, our affinity based method has the unique merits of avoiding per-domain hard clustering and having fewer parameters to tune (e.g. the per-camera cluster number).

3 Method Formulation

To overcome the limitation of supervised model learning algorithms for exhaustive within-camera and cross-camera ID labelling, we propose a novel Unsupervised Tracklet Association Learning (UTAL) method to person re-id in videos (or multi-shot images in general). This is achieved by exploiting person *tracklet labelling* obtained from existing trackers[2] *without* any ID labelling either cross-camera or within-camera. The UTAL learns a person re-id model end-to-end therefore benefiting from joint overall model optimisation in deep learning. In the follows, we first present unsupervised per-camera tracklet labelling (Sec. 3.1), then describe our model design for within-camera and cross-camera tracklet association by joint unsupervised deep learning (Sec. 3.2).

3.1 Unsupervised Per-Camera Tracklet Formation

Given a large quantity of video data captured by disjoint surveillance cameras, we first deploy the off-the-shelf pedestrian detection and tracking models [50,51,52] to automatically extract person tracklets. We then annotate each tracklet $\boldsymbol{S}$ with a unique class (one-hot) label y in an *unsupervised* and *camera-independent* manner. This does not involve any *manual* ID verification on tracklets. By applying this tracklet labelling method in each camera view separately, we can obtain an *independent* set of labelled tracklets $\{\boldsymbol{S}_i, y_i\}$ per camera, where each tracklet $\boldsymbol{S}$ contains a varying number of person bounding box images $\boldsymbol{I}$ as $\boldsymbol{S} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, \cdots\}$.

Challenges. To effectively learn a person re-id model from such automatically labelled tracklet training data, we need to deal with two modelling challenges centred around the supervision of person ID class labels: (1) Due to frequent trajectory fragmentation, multiple tracklets (unknown due to no manual verification) are often generated during the appearing period of a person under one camera view. However, they are unsupervisedly assigned with different one-hot categorical labels. This may significantly mislead the discriminative learning process of a re-id model. (2) There are no access to positive and negative pairwise ID correspondence between tracklet labels across disjoint camera views. Lacking cross-camera person ID supervision underpins one of the key challenges in unsupervised tracklet person re-id.

3.2 Unsupervised Tracklet Association

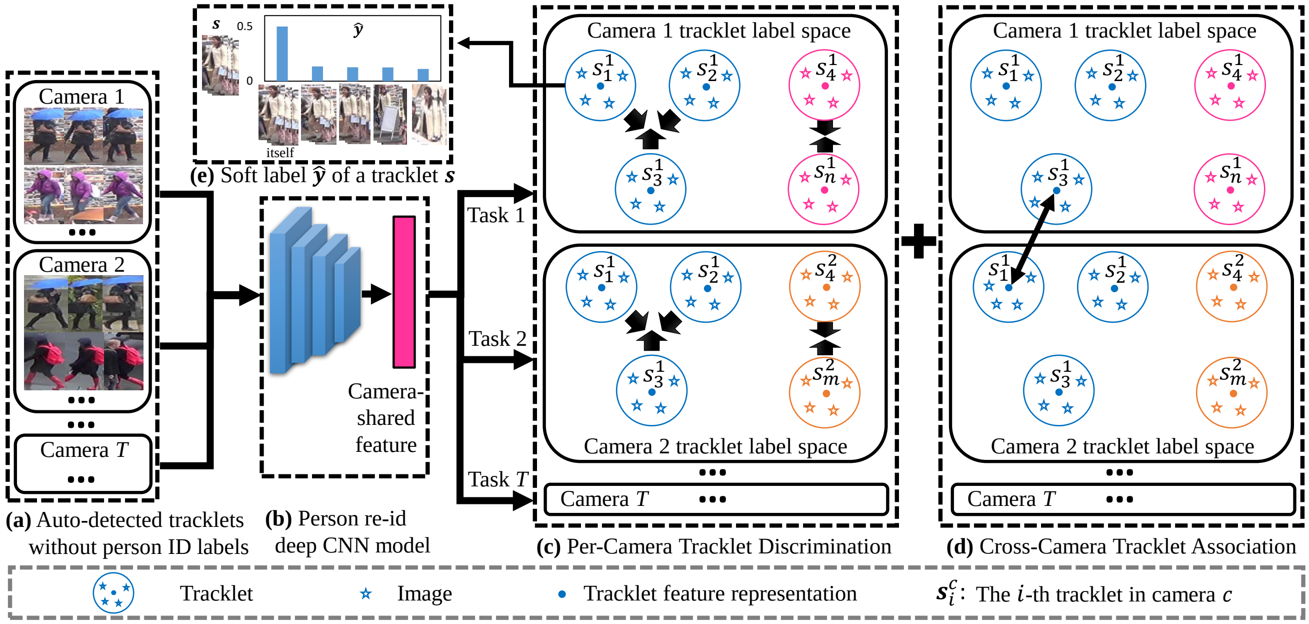
Given per-camera independent tracklets $\{\boldsymbol{S}_i, y_i\}$, we explore *tracklet label re-id discriminative learning* without person ID labels in a deep learning classification framework. We formulate an ***Unsupervised Tracklet Association Learning*** (UTAL) method, with the overall architecture design illustrated in Fig. 1. The UTAL contains two model components: **(I)** *Per-Camera Tracklet Discrimination* (PCTD) learning for optimising "local" within-camera tracklet label discrimination. This facilitates "global" cross-camera tracklet association, given *independent* tracklet label spaces in different camera views. **(II)** *Cross-Camera Tracklet Association* (CCTA) learning for discovering "global" cross-camera tracklet identity matching without ID labels.

For accurate cross-camera tracklet association, it is important to formulate a robust image feature representation to characterise the person appearance of each tracklet. However, it is sub-optimal to achieve "local" per-camera tracklet discriminative learning using only per-camera independent tracklet labels without "global" cross-camera tracklet correlations. We therefore propose to optimise jointly both PCTD and CCTA. The two components integrate as a whole in a single deep learning architecture, learn jointly and mutually benefit each other in incremental end-to-end model optimisation. Our overall idea for unsupervised learning of tracklet person re-id is to maximise coarse-grained *latent group-level* cross-camera tracklet association. This is based on exploring an notion of tracklet set correlation learning (Fig. 2(b)). It differs significantly from supervised re-id learning that relies heavily on the fine-grained *explicit instance-level* cross-camera ID pairwise supervision (Fig. 2(a)).

3.2.1 Per-Camera Tracklet Discrimination Learning

In PCTD learning, we treat each individual camera view separately. That is, optimising per-camera labelled tracklet discrimination as a classification task with the unsupervised per-camera tracklet labels (not person ID labels) (Fig. 1(a)). Given a surveillance network with T cameras, we hence have a total of T different tracklet classification tasks each corresponding to a specific camera view.

[2] Although object tracklets can be generated by any independent single-camera-view multi-object tracking (MOT) models widely available currently, a conventional MOT model is *not* end-to-end optimised for cross-camera tracklet association.


Camera 1
Camera 2
Camera T
s
0.5
0
ŷ
itself
Camera-shared feature
Task 1
Task 2
Task T
Camera 1 tracklet label space
Camera 2 tracklet label space
Camera T
Tracklet
Image
Tracklet feature representation
s_i^c: The i-th tracklet in camera c

(a) Auto-detected tracklets without person ID labels **(b)** Person re-id deep CNN model **(c)** Per-Camera Tracklet Discrimination **(d)** Cross-Camera Tracklet Association **(e)** Soft label $\hat{\boldsymbol{y}}$ of a tracklet $\boldsymbol{s}$

Fig. 1. An overview of the proposed *Unsupervised Tracklet Association Learning* (UTAL) person re-identification model. The UTAL takes as input **(a)** auto-detected tracklets from all the camera views without any person ID class labelling either within-camera or cross-camera. The objective is to derive **(b)** a person re-id discriminative feature representation model by unsupervised learning. To this end, we formulate the UTAL model for simultaneous **(c)** Per-Camera Tracklet Discrimination (PCTD) learning and **(d)** Cross-Camera Tracklet Association (CCTA) learning in an end-to-end neural network architecture. The PCTD aims to derive the "local" discrimination of per-camera tracklets in the respective tracklet label space (represented by soft labels **(e)**) by a multi-task inference process (one task for a specific camera view), whilst the CCTA to learn the "global" cross-camera tracklet association across independently formed tracklet label spaces. In UTAL design, the PCTD and CCTA jointly learn to optimise a re-id model for maximising their complementary contributions and advantages in a synergistic interaction and integration. Best viewed in colour.

Importantly, we further formulate these T classification tasks in a multi-branch network architecture design. All the tasks share the *same* feature representation space (Fig. 1(b)) whilst enjoying an individual classification branch (Fig. 1(c)). This is a multi-task learning [42].

Formally, we assume M_t different tracklet labels $\{y\}$ with the training tracklet image frames $\{\boldsymbol{I}\}$ from a camera view $t \in \{1, \cdots, T\}$ (Sec. 3.1). We adopt the softmax Cross-Entropy (CE) loss function to optimise the corresponding classification task (the t-th branch). The CE loss on a training image frame $(\boldsymbol{I}, y)$ is computed as:

$$\mathcal{L}_{\text{ce}} = -\sum_{j=1}^{M_t} \mathbb{1}(j = y) \cdot \log\Big(\frac{\exp(\boldsymbol{W}_j^\top \boldsymbol{x})}{\sum_{k=1}^{M_t} \exp(\boldsymbol{W}_k^\top \boldsymbol{x})}\Big) \tag{1}$$

where $\boldsymbol{x}$ specifies the feature vector of $\boldsymbol{I}$ extracted from the *task-shared* representation space and $\boldsymbol{W}_y$ the y-th class prediction parameters. $\mathbb{1}(\cdot)$ denotes an indicator function that returns $1/0$ for true/false arguments. Given a training mini-batch, we compute the CE loss for each such training sample with the respective tracklet label space and utilise their average to form the PCTD learning objective as:

$$\mathcal{L}_{\text{pctd}} = \frac{1}{N_{\text{bs}}} \sum_{t=1}^{T} \mathcal{L}_{\text{ce}}^t \tag{2}$$

where $\mathcal{L}_{\text{ce}}^t$ denotes the CE loss of all training tracklet frames from the t-th camera, and N_{bs} specifies the batch size.

Recall that, one of the main challenges in unsupervised tracklet re-id learning arises from within-camera trajectory fragmentation. This causes the tracklet ID duplication issue, i.e. the same-ID tracklets are assigned with distinct labels. By treating every single tracklet label as a unique class (Eq (2)), misleading supervision can be resulted potentially hampering the model learning performance.

Soft Labelling. For gaining learning robustness against unconstrained trajectory fragmentation, we exploit the pairwise appearance affinity (similarity) information between within-camera tracklets. To this end, we propose *soft tracklet labels* to replace the *hard* counterpart (one-hot labels). This scheme uniquely takes into account the underlying ID correlation between tracklets in the PCTD learning (Eq (2)). It is based on the intuition that, multiple fragmented tracklets of the same person are more likely to share higher visual affinity with each other than those describing different people. Therefore, using tracklet labels involving the appearance affinity (i.e. soft labels) means imposing person ID relevant information into model training, from the manifold structure learning perspective [53].

Formally, we start the computation of soft tracklet labels by constructing an affinity matrix of person appearance $\mathcal{A}^t \in \mathbb{R}^{M_t \times M_t}$ on all M_t tracklets for each camera $t \in \{1, \cdots, T\}$, where each element $\mathcal{A}^t(i,j)$ specifies the visual appearance similarity between the tracklets i and j. This requires a tracklet feature representation space. We achieve this by cumulatively updating an external feature vector $\boldsymbol{s}$ for every single tracklet $\boldsymbol{S}$ with the image features of the constituent frames in a batch-wise manner.

More specifically, given a mini-batch including n_i^t image frames from the i-th tracklet $\boldsymbol{S}_i^t$ of t-th camera view, the corresponding tracklet representation $\boldsymbol{s}_i^t$ is progressively updated across the training iterations as:

$$\boldsymbol{s}_i^t = \frac{1}{1+\alpha}\Big[\boldsymbol{s}_i^t + \alpha\Big(\frac{1}{n_i^t}\sum_{k=1}^{n_i^t} \boldsymbol{x}_k\Big)\Big] \tag{3}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2903058, IEEE Transactions on Pattern Analysis and Machine Intelligence

where $\boldsymbol{x}_k$ is the feature vector of the k-th in-batch image frame of $\boldsymbol{S}_i^t$, extracted by the up-to-date model. The learning rate parameter α controls how fast $\boldsymbol{s}_i^t$ updates. This method avoids the need of forwarding all tracklet data in each iteration therefore computationally efficient and scalable.

Given the tracklet feature representations, we subsequently sparsify the affinity matrix as:

$$\mathcal{A}^t(i,j) = \begin{cases} \exp(-\frac{\|\boldsymbol{s}_i^t - \boldsymbol{s}_j^t\|_2^2}{\sigma^2}), & \text{if } \boldsymbol{s}_j^t \in \mathcal{N}(\boldsymbol{s}_i^t) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\mathcal{N}(\boldsymbol{s}_i^t)$ are the K nearest neighbours (NN) of $\boldsymbol{s}_i^t$ defined by the Euclidean distance in the feature space. Using the sparse NN idea on the affinity structure is for suppressing the distracting effect of visually similar tracklets from unmatched ID classes. Computationally, each $\mathcal{A}$ has a quadratic complexity, but only to the number of *per-camera* tracklet and linear to the total number of cameras, rather than quadratic to all tracklets from all the cameras. The use of sparse similarity matrices significantly reduces the memory demand.

To incorporate the local density structure [54], we deploy a neighbourhood structure-aware scale defined as:

$$\sigma^2 = \frac{1}{M_t \cdot K} \sum_{i=1}^{M_t} \sum_{j=1}^{K} \|\boldsymbol{s}_i^t - \boldsymbol{s}_j^t\|_2^2, \;\; s.t. \;\; \boldsymbol{s}_j^t \in \mathcal{N}(\boldsymbol{s}_i^t) \tag{5}$$

Based on the estimated neighbourhood structures, we finally compute the soft label (Fig. 1(e)) for each tracklet $\boldsymbol{S}_i^t$ as the L_1 normalised affinity measurement:

$$\hat{\boldsymbol{y}}_i^t = \frac{\mathcal{A}(i, 1:M_t)}{\sum_{j=1}^{M_t}\Big(\mathcal{A}(i,j)\Big)} \tag{6}$$

Given the proposed soft tracklet labels, the CE loss function (Eq (2)) is then reformulated as:

$$\mathcal{L}_{\text{sce}} = -\sum_{j=1}^{M_t} \hat{\boldsymbol{y}}_i^t(j) \cdot \log\Big(\frac{\exp(\boldsymbol{W}_j^\top \boldsymbol{x})}{\sum_{k=1}^{M_t}\exp(\boldsymbol{W}_k^\top \boldsymbol{x})}\Big) \tag{7}$$

We accordingly update the PCTD learning loss (Eq (2)) as:

$$\mathcal{L}_{\text{pctd}} = \frac{1}{N_{\text{bs}}} \sum_{t=1}^{T} \mathcal{L}_{\text{sce}}^t. \tag{8}$$

Remarks. In PCTD, the objective function (Eq (8)) optimises by supervised learning person tracklet discrimination *within* each camera view alone. It does not explicitly consider supervision in *cross-camera* tracklet association. Interestingly, when jointly learning all the per-camera tracklet discrimination tasks together, the learned representation model is *implicitly* and *collectively* cross-view tracklet discriminative in a latent manner. This is due to the existence of cross-camera tracklet ID class correlation. That being said, the shared feature representation is optimised to be tracklet discriminative *concurrently* for all camera views, latently expanding model discriminative learning from per-camera (*locally*) to cross-camera (*globally*).

Apart from multi-camera multi-task learning, we exploit the idea of soft tracklet labels to further improve the model ID discrimination learning capability. This is for better robustness against trajectory fragmentation. Fundamentally, this is an indirect strategy of refining fragmented tracklets.

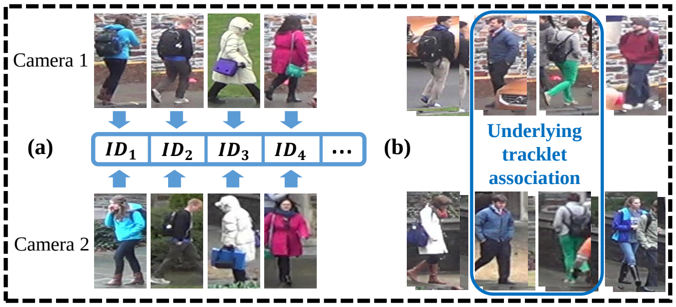


Fig. 2. Comparing **(a)** Fine-grained *explicit instance-level* cross-camera ID labelled image pairs for supervised person re-id model learning and **(b)** Coarse-grained *latent group-level* cross-camera tracklet (a multi-shot group) label correlation for ID label-free (unsupervised) person re-id learning using the proposed UTAL method.

It is based on the visual appearance affinity without the need of explicitly stitching tracklets. The intuition is that, the tracklets of the same person are possible to be assigned with more similar soft labels (i.e. signatures). Consequently, this renders the unsupervised tracklet labels closer to supervised ID class labels in terms of discrimination power, therefore helping re-id model optimisation.

In equation formulation, our PCTD objective is related to the Knowledge Distillation (KD) technique [55]. KD also utilises soft class probability labels inferred by an independent teacher model. Nevertheless, our method conceptually differs from KD, since we primarily aim to unveil the hidden fine-grained discriminative information of the same class (ID) distributed across unconstrained tracklet fragments, Besides, our model retains the KD's merit of modelling the inter-class similarity geometric manifold information. Also, our method has no need for learning a heavy source knowledge teacher model, therefore, computationally more efficient. We will evaluate the PCTD model design (Table 4).

3.2.2 *Cross-Camera Tracklet Association Learning*

The PCTD achieves somewhat global (all the camera views) tracklet discrimination capability *implicitly*. But the resulting representation remains sub-optimal, due to the lack of *explicitly* optimising cross-camera tracklet association at the fine-grained instance level. It is non-trivial to impose cross-camera re-id discriminative learning constraints at the absence of ID labels. To address this problem, we introduce a Cross-Camera Tracklet Association (CCTA) learning algorithm for enabling tracklet association between cameras (Fig. 1(d)). Conceptually, the CCTA is based on *adaptively and incrementally self-discovering cross-view tracklet association* in the multi-task camera-shared feature space (Fig. 1(b)).

Specifically, we ground the CCTA learning on cross-camera nearest neighbourhoods. In re-id, the vast majority of cross-camera tracklet pairs are negative associations from unmatched ID classes. They provide no desired information about how a person's appearance varies under different camera viewing conditions. The key for designing an informative CCTA loss is therefore to self-discover the cross-camera positive *matching* pairs. This requires to search similar samples (i.e. neighbours) which however is a challenging task because: (1) Tracklet feature representations $\boldsymbol{s}$ can be unreliable and error-prone due to the lack of cross-camera pair supervision (hence a catch-22 problem). (2) False positive pairs may easily propagate the erroneous supervision cumulatively through the learning process, guiding the optimisation towards poor local optima; Deep neural networks possess the capacity to fit any supervision labels [56].

To overcome these challenges, we introduce a model matureness adaptive matching pair search mechanism. It progressively finds an *increasing* number of plausible true matches across cameras as a reliable basis for the CCTA loss formulation. In particular, in each training epoch, we first retrieve the reciprocal cross-camera nearest neighbour $\mathcal{R}(\boldsymbol{s}_i^t)$ for each tracklet $\boldsymbol{s}_i^t$. The $\mathcal{R}$ is obtained based on the mutual nearest neighbour notion [57]. Formally, let $\mathcal{N}^1(\boldsymbol{s}_i^t)$ be the cross-camera NN of $\boldsymbol{s}_i^t$. The $\mathcal{R}(\boldsymbol{s}_i^t)$ is then defined as:

$$\mathcal{R}(\boldsymbol{s}_i^t) = \{\boldsymbol{s} | \boldsymbol{s} \in \mathcal{N}^1(\boldsymbol{s}_i^t) \ \ \&\& \ \ \boldsymbol{s}_i^t \in \mathcal{N}^1(\boldsymbol{s})\} \tag{9}$$

Given such self-discovered cross-camera matching pairs, we then formulate a CCTA objective loss for a tracklet $\boldsymbol{s}_i^t$ as:

$$\mathcal{L}_{\text{ccta}} = \sum_{\boldsymbol{s} \in \mathcal{R}(\boldsymbol{s}_i^t)} \parallel \boldsymbol{s}_i^t - \boldsymbol{s} \parallel_2 \tag{10}$$

With Eq (10), we impose a cross-camera discriminative learning constraint by encouraging the model to pull the neighbour tracklets in $\mathcal{R}_i^t$ close to $\boldsymbol{s}_i^t$. This CCTA loss applies only to those tracklets $\boldsymbol{s}$ with cross-camera matches, i.e. $\mathcal{R}(\boldsymbol{s})$ is non-empty, so that it is model matureness adaptive. As the training proceeds, the model is supposed to become more mature, leading to more cross-camera tracklet matches discovered. We will evaluate the CCTA loss in Sec. 4.3.

Remarks. Under the mutual nearest neighbour constraint, $\mathcal{R}(\boldsymbol{s}_i^t)$ are considered to be more strictly similar to $\boldsymbol{s}_i^t$ than the conventional nearest neighbours $\mathcal{N}(\boldsymbol{s}_i^t)$. With uncontrolled viewing condition variations across cameras, matching tracklets with dramatic appearance changes may be excluded from the $\mathcal{R}(\boldsymbol{s}_i^t)$ particularly in the beginning, representing a conservative search strategy. This is designed so to minimise the negative effect of error propagation from false matching pairs. More matching pairs are likely unveiled as the training proceeds. Many previously missing pairs can be gradually discovered when the model becomes more mature and discriminative. Intuitively, easy matching pairs are found before hard ones. Hence, the CCTA loss is in a curriculum leaning spirit [58]. In Eq (10) we consider only the positive pairs whilst ignoring the negative matches. This is conceptually analogue to the formulation of Canonical Correlation Analysis (CCA) [59], and results in a simpler objective function without the need to tune a margin hyper-parameter as required by the ranking losses [60].

3.2.3 Joint Unsupervised Tracklet Association Learning

By combining the CCTA and PCTD learning constraints, we obtain the final model objective loss function of UTAL as:

$$\mathcal{L}_{\text{utal}} = \mathcal{L}_{\text{pctd}} + \lambda \mathcal{L}_{\text{ccta}} \tag{11}$$

where λ is a balance weight. Note that $\mathcal{L}_{\text{pctd}}$ is an average loss term at the individual tracklet image level whilst $\mathcal{L}_{\text{ccta}}$ at the tracklet group (set) level. Both are derived from the same mini-batch of training data concurrently.

Remarks. By design, the CCTA enhances model representation learning. It imposes discriminative constraints derived from self-discovered cross-camera tracklet association. This is based on the PCTD learning of unsupervised and per-camera independent tracklet label spaces. With more discriminative representation in the subsequent training iterations, the PCTD is then able to deploy more accurate soft tracklet labels. This in turn facilitates not only the following representation learning of per-camera tracklets, but also the discovery of higher quality and more informative cross-camera tracklet matching pairs. In doing so, the two learning components integrate seamlessly and optimise a person re-id model concurrently in an end-to-end batch-wise learning process. Consequently, the overall UTAL method formulates a benefit-each-other closed-loop model design. This eventually leads to cumulative and complementary advantages throughout training.

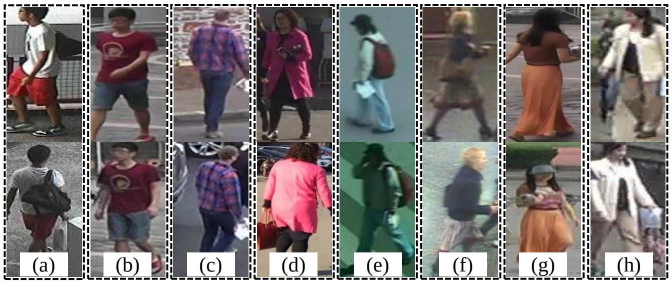


Fig. 3. Example cross-camera matching image/tracklet pairs from (a) CUHK03, (b) Market-1501, (c) DukeMTMC-ReID, (d) MSMT17, (e) PRID2011, (f) iLIDS-VID, (g) MARS, (h) DukeMTMC-SI-Tracklet.

3.2.4 Model Training and Testing

Model Training. To minimise the negative effect of inaccurate cross-camera tracklet matching pairs, we deploy the CCTA loss *only* during the second half training process. Specifically, UTAL begins the model training with the soft tracklet label based PCTD loss (Eq (8)) for the first half epochs. We then deploy the full UTAL loss (Eq (11)) for the remaining epochs. To improve the training efficiency, we update the per-camera tracklet soft labels (Eq (6)) and cross-camera tracklet matches (Eq (9)) per epoch. These updates progressively enhance the re-id discrimination power of the UTAL objective throughout training, as we will show in our model component analysis and diagnosis in Sec. 4.3.

Model Testing. Once a deep person re-id model is trained by the UTAL unsupervised learning method, we deploy the camera-shared feature representations (Fig. 1(b)) for re-id matching under the Euclidean distance metric.

4 Experiments

4.1 Experimental Setting

Datasets. To evaluate the proposed UTAL model, we tested both video (iLIDS-VID [25], PRID2011 [26], MARS [27]) and image (CUHK03 [21], Market-1501 [22], DukeMTMC-ReID [23,24], MSMT17 [4]) person re-id datasets. In previous studies, these two sets of benchmarks were usually evaluated *separately*. We consider both sets because recent large image re-id datasets were typically constructed by sampling person bounding boxes from videos, so they share similar characteristics as the video re-id datasets. We adopted the standard test protocols as summarised in Table 1.

To further test realistic model performances, we introduced a new tracklet person re-id benchmark based on

TABLE 1
Dataset statistics and evaluation setting.

Dataset	# ID	# Train	# Test	# Image	# Tracklet
iLIDS-VID [25]	300	150	150	43,800	600
PRID2011 [26]	178	89	89	38,466	354
MARS [27]	1,261	625	636	1,191,003	20,478
DukeMTMC-SI-Tracklet	1,788	702	1,086	833,984	12,647
CUHK03 [21]	1,467	767	700	14,097	0
Market-1501 [22]	1,501	751	750	32,668	0
DukeMTMC-ReID [24]	1,812	702	1,110	36,411	0
MSMT17 [4]	4,101	1,041	3,060	126,441	0

DukeMTMC [23]. It differs from all the existing DukeMTMC variants [24,66,67] by uniquely providing *automatically* generated tracklets. We built this new tracklet person re-id dataset as follows. We first deployed an efficient deep learning tracker that leverages a COCO+PASCAL trained *SSD* [50] for pedestrian detection and an ImageNet trained *Inception* [68] for person appearance matching. Applying this tracker to all DukeMTMC raw videos, we generated 19,135 person tracklets. Due to the inevitable detection and tracking errors caused by background clutters and visual ambiguity, these tracklets may present typical mistakes (e.g. ID switch) and corruptions (e.g. occlusion). We name this test ***DukeMTMC-SI-Tracklet***, abbreviated as ***DukeTracklet***.

For benchmarking DukeTracklet, we need the ground-truth person ID labels of tracklets. To this end, we used the criterion of spatio-temporal average Intersection over Union (IoU) between detected tracklets and ground-truth trajectories available in DukeMTMC. In particular, we labelled an auto-generated tracklet by the ground-truth person ID associated with a manually-generated trajectory if their average IoU is over 50%. Otherwise, we labelled the auto-generated tracklet as "unknown ID". To maximise the comparability with existing DukeMTMC variants, we threw away those tracklets labelled with unknown IDs. We finally obtained 12,647 person tracklets from 1,788 unique IDs. The average tracklet duration is 65.9 frames. To match DukeMTMC-ReID [24], we set the same 702 training IDs with the remaining 1,086 people for performance test (missing 14 test IDs against DukeMTMC-ReID due to tracking failures).

Tracklet Label Assignment. For each video re-id dataset, we simply assigned each tracklet with a unique label in a camera-independent manner (Sec. 3.1). For each multi-shot image datasets, we assumed all person images per ID per camera were drawn from a single pedestrian tracklet, and similarly labelled them as the video datasets.

Performance Metrics. We adopted the common Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) metrics [22] for model performance measurement.

Implementation Details. We used an ImageNet pre-trained ResNet-50 [69] as the backbone net for UTAL, along with an additional 2,048-D fully-connected (FC) layer for deriving the camera-shared representations. Every camera-specific branch was formed by one FC classification layer. Person bounding box images were resized to 256×128. To ensure each training mini-batch has person images from all cameras, we set the batch size to 128 for PRID2011, iLIDS-VID and CUHK03, and 384 for MSMT17, Market-1501, MARS, DukeMTMC-ReID, and DukeTracklet. In order to balance the model training speed across cameras, we randomly selected the same number of tracklets per camera and the same number of frame images (4 images) per chosen tracklet when sampling each mini-batch. We adopted the Adam optimiser [70] with the learning rate of 3.5×10^{-4} and the epoch of 200. By default, we set $\lambda = 10$ for Eq (11), $\alpha = 1$ for Eq (3), and $K = 4$ for Eq (5) in the following experiments.

4.2 Comparisons to the State-Of-The-Art Methods

We compared two different sets of state-of-the-art methods on image and video re-id datasets, due to the independent studies on them in the literature.

Evaluation on Image Datasets. Table 2 shows the unsupervised re-id performance of the proposed UTAL and 15 state-of-the-art methods including 3 hand-crafted feature based methods (Dic [9], ISR [10], RKSL [13]) and 12 auxiliary knowledge (identity/attribute) transfer based models (AE [61], AML [63], UsNCA [64], CAMEL [20], JSTL [62], PUL [19], TJ-AIDL [17], CycleGAN [35], SPGAN [36], HHL [37], DASy [40]). The results show four observations as follows.

(1) Among the existing methods, the knowledge transfer based models are often superior due to the use of *additional* label information, e.g. Rank-1 39.4% by CAMEL *vs* 36.5% by Dic on CUHK03; 65.7% by DASy *vs* 50.2% by Dic on Market-1501. To that end, CAMEL needs to benefit from learning on 7 different person re-id datasets of diverse domains (CUHK03 [21], CUHK01 [74], PRID [26], VIPeR [75], i-LIDS [76]) including a total of 44,685 images and 3,791 IDs; HHL requires to utilise labelled Market-1501 (750 IDs) or DukeMTMC-ReID (702 IDs) as the source training data. DASy needs elaborative ID synthesis and adaptation.

(2) The proposed UTAL outperforms all competitors with significant margins. For example, the Rank-1 margin by UTAL over HHL is 7.0% (69.2-62.2) on Market-1501 and 15.4% (62.3-46.9) on DukeMTMC-ReID. Also, our preliminary method TAUDL already surpasses all previous methods. It is worth pointing out that UTAL dose not benefit from any additional labelled source domain training data as compared to the strong alternative HHL. Importantly, UTAL is potentially more scalable due to *no* reliance at all on the similarity constraint between source and target domains.

(3) The UTAL is simpler to train with a simple end-to-end model learning, *vs* the alternated deep CNN training and data clustering required by PUL, a two-stage model training of TJ-AIDL, high GAN training difficulty of HHL, and elaborative ID synthesis of DASy. These results show both the performance advantage and model design superiority of UTAL over state-of-the-art re-id methods.

(4) A large performance gap exists between unsupervised and supervised learning models. Further improvement is required on unsupervised learning algorithms.

Evaluation on Video Datasets. In Table 3, we compared the UTAL with 8 state-of-the-art unsupervised video re-id models (GRDL [11], UnKISS [12], SMP [16], DGM+MLAPG/IDE [15], DAL [72], RACE [71], DASy [40]) on the video benchmarks. Unlike UTAL, all these existing models except DAL are *not* end-to-end deep learning methods using hand-crafted or independently trained deep features as input.

The comparisons show that, our UTAL outperforms all existing video person re-id models on the large scale video dataset MARS, e.g. by a Rank-1 margin of 3.1% (49.9-46.8) and a mAP margin of 13.8% (35.2-21.4) over the best competitor DAL. However, UTAL is inferior than top existing

TABLE 2
Unsupervised person re-id on image based datasets. *: Benefited from extra labelled auxiliary training data. "-": No reported result.

Dataset	CUHK03 [21]		Market-1501 [22]		DukeMTMC-ReID [24]		MSMT17 [4]	
Metric (%)	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Dic [9]	36.5	-	50.2	22.7	-	-	-	-
ISR [10]	38.5	-	40.3	14.3	-	-	-	-
RKSL [13]	34.8	-	34.0	11.0	-	-	-	-
SAE* [61]	30.5	-	42.4	16.2	-	-	-	-
JSTL* [62]	33.2	-	44.7	18.4	-	-	-	-
AML* [63]	31.4	-	44.7	18.4	-	-	-	-
UsNCA* [64]	29.6	-	45.2	18.9	-	-	-	-
CAMEL* [20]	39.4	-	54.5	26.3	-	-	-	-
PUL* [19]	-	-	44.7	20.1	30.4	16.4	-	-
TJ-AIDL* [17]	-	-	58.2	26.5	44.3	23.0	-	-
CycleGAN* [35]	-	-	48.1	20.7	38.5	19.9	-	-
SPGAN* [36]	-	-	51.5	22.8	41.1	22.3	-	-
SPGAN+LMP* [36]	-	-	57.7	26.7	46.4	26.2	-	-
HHL* [37]	-	-	62.2	31.4	46.9	27.2	-	-
DASy* [40]	-	-	65.7	-	-	-	-	-
TAUDL [28]	44.7	31.2	63.7	41.2	61.7	43.5	28.4	12.5
UTAL	**56.3**	**42.3**	**69.2**	**46.2**	**62.3**	**44.6**	**31.4**	**13.1**
GCS [65](*Supervised*)	88.8	97.2	93.5	81.6	84.9	69.5	-	-

TABLE 3
Unsupervised person re-id on video based datasets. *: Assume no tracking fragmentation. †: Use some ID labels for model initialisation.

Dataset	PRID2011 [26]			iLIDS-VID [25]			MARS [27]				DukeTracklet			
Metric (%)	Rank-1	Rank-5	Rank-20	Rank-1	Rank-5	Rank-20	Rank-1	Rank-5	Rank-20	mAP	Rank-1	Rank-5	Rank-20	mAP
GRDL [11]	41.6	76.4	89.9	25.7	49.9	77.6	19.3	33.2	46.5	9.6	-	-	-	-
UnKISS [12]	58.1	81.9	96.0	35.9	63.3	83.4	22.3	37.4	53.6	10.6	-	-	-	-
SMP* [16]	80.9	95.6	99.4	41.7	66.3	80.7	23.9	35.8	44.9	10.5	-	-	-	-
DGM+MLAPG† [15]	73.1	92.5	99.0	37.1	61.3	82.0	24.6	42.6	57.2	11.8	-	-	-	-
DGM+IDE† [15]	56.4	81.3	96.4	36.2	62.8	82.7	36.8	54.0	68.5	21.3	-	-	-	-
RACE† [71]	50.6	79.4	91.8	19.3	39.3	68.7	43.2	57.1	67.6	24.5	-	-	-	-
DASy* [40]	43.0	-	-	56.5	-	-	-	-	-	-	-	-	-	-
DAL [72]	**85.3**	**97.0**	**99.6**	**56.9**	**80.6**	**91.9**	46.8	63.9	77.5	21.4	-	-	-	-
TAUDL [28]	49.4	78.7	98.9	26.7	51.3	82.0	43.8	59.9	72.8	29.1	26.1	42.0	57.2	20.8
UTAL	54.7	83.1	96.2	35.1	59.0	83.8	**49.9**	**66.4**	**77.8**	**35.2**	**43.8**	**62.8**	**76.5**	**36.6**
Snippet [73](*Supervised*)	93.0	99.3	100.0	85.4	96.7	99.5	86.3	94.7	98.2	76.1	-	-	-	-

models on the two small benchmarks iLIDS-VID (300 training tracklets) and PRID2011 (178 training tracklets), *vs* 8,298 training tracklets in MARS. This shows that UTAL does need sufficient tracklet data in order to have its performance advantage. As the required tracklet data are not manually labelled, this requirement is not a hindrance to its scalability on large scale deployments. Quite the contrary, UTAL works the best when large unlabelled video data are available. A model would benefit from pre-training using UTAL on large auxiliary unlabelled videos with similar viewing conditions.

4.3 Component Analysis and Discussion

We conducted detailed UTAL model component analysis on two large tracklet re-id datasets, MARS and DukeTracklet.

Per-Camera Tracklet Discrimination Learning. We started by testing the performance impact of the PCTD component. This is achieved by designing a baseline that treats all cameras together, that is, concatenating the per-camera tracklet label spaces and deploying the Cross-Entropy loss for learning a ***Single-Task Classification*** (STC). In this analysis, we did not consider the cross-camera tracklet association component for a more focused evaluation.

Table 4 shows that, the proposed PCTD design is significantly superior over the STC learning algorithm, e.g. achieving Rank-1 gain of 27.9% (43.8-15.9), and 27.8% (31.7-3.9) on MARS and DukeTracklet, respectively. The results demonstrate modelling advantages of PCTD in exploiting unsupervised tracklet labels for learning cross-view re-id discriminative features. This validates the proposed idea of implicitly deriving a cross-camera shared feature representation through a multi-camera multi-task learning strategy.

TABLE 4
Effect of Per-Camera Tracklet Discrimination (PCTD) learning.

Dataset	MARS [27]		DukeTracklet	
Metric (%)	Rank-1	mAP	Rank-1	mAP
STC	15.9	10.0	3.9	4.7
PCTD	**43.8**	**31.4**	**31.7**	**26.4**

Recall that we propose a soft label (Eq (6)) based Cross-Entropy loss (Eq (7)) for tackling the notorious trajectory fragmentation challenge. To test how much benefit our soft labelling strategy brings to unsupervised tracklet re-id, we compared it against the one-hot class *hard*-labelling counterpart (Eq (1)). Table 5 shows that the proposed soft-labelling is significantly superior, suggesting a clear benefit in mitigating the negative impact of trajectory fragmentation. This is due to the intrinsic capability of exploiting the appearance pairwise affinity knowledge among tracklets per camera.

TABLE 5
Soft-labelling *versus* hard-labelling.

Dataset	MARS [27]		DukeTracklet	
Metric (%)	Rank-1	mAP	Rank-1	mAP
Hard-Labelling	35.5	20.5	24.5	17.3
Soft-Labelling	**49.9**	**35.2**	**43.8**	**36.6**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2903058, IEEE Transactions on Pattern Analysis and Machine Intelligence

We further examined the ID discrimination capability of tracklet soft labels that underpins its outperforming over the corresponding hard labels. To this end, we measured the *Mean Pairwise Similarity* (MPS) of soft label vectors assigned to per-camera same-ID tracklets. Figure 4 shows that, the MPS metric goes higher as the training epoch increases, particularly after the CCTA loss is further exploited in the middle of training (at 100th epoch). This indicates explicitly the evolving process of mining the discriminative knowledge among fragmented tracklets with the same person ID labels in a self-supervising fashion.

In effect, the affinity measurement (Eq (4)) used for computing soft labels can be useful for automatically merging short fragmented tracklets into long trajectories per camera. We tested this *tracking refinement* capability of our method. In particular, we built a sparse connection graph by thresholding the pairwise affinity scores at 0.5, analysed the connected tracklet components [77], and merged all tracklets in each component into a long trajectory.

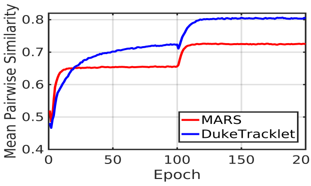


Fig. 4. The evolving process of tracklet soft label quality over the model training epochs on MARS and DukeTracklet.

Table 6 shows that with our per-camera tracklet affinity measurement, even such a simple strategy can merge 4,389/2,527 out of 8,298/5,803 short tracklets into 1,532/982 long trajectories at the NMI (Normalised Mutual Information) rate of 0.896/0.934 on MARS/DukeTracklet. This not only suggests the usefulness of UTAL in tracklet refinement, but also reflects the underlying correlation between tracking and person re-id. For visual quality examination, we gave example cases for tracking refinement in Fig. 5.

TABLE 6
Evaluating the tracking refinement capability of soft labels.

Dataset	MARS [27]	DukeTracklet
Original Tracklets	8,298	5,803
Mergable Tracklets	4,389	2,527
Long Trajectories	1,532	928
NMI	0.896	0.934

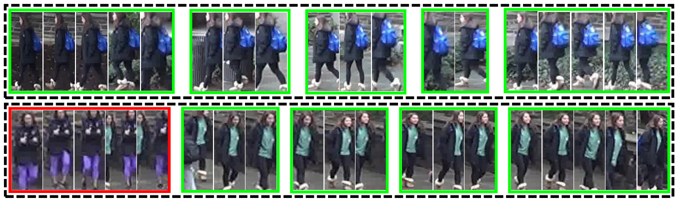


Fig. 5. Example long trajectories discovered by UTAL among unlabelled short fragmented tracklets. Each row denotes a case. The tracklets in green/red bounding box denote the true/false matches, respectively. Failure tracklet merging may be due to detection and tracking errors.

Algorithmically, our soft label PCTD naturally inherits the cross-class (ID) knowledge transfer capability from *Knowledge Distillation* [55]. It is interesting to see how much performance benefit this can bring to unsupervised tracklet re-id. To this end, we conducted a controlled experiment with only one randomly selected training tracklet per ID per camera. Doing so enforces that no multiple per-camera tracklets share the same person ID, which more explicitly evaluates the impact of cross-ID knowledge transfer. Note, this leads to probably inferior re-id model generalisation capability due to less training data used in optimisation. Table 7 shows that the cross-ID knowledge transfer gives notable performance improvements.

TABLE 7
Evaluating the cross-ID knowledge transfer effect of soft labels.

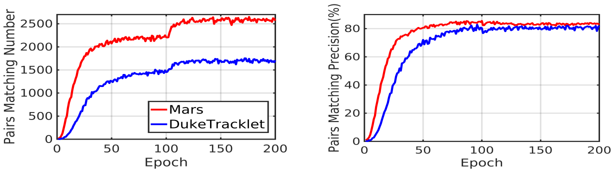
Dataset	MARS [27]		DukeTracklet	
Metric (%)	Rank-1	mAP	Rank-1	mAP
Hard Label	45.1	31.1	28.6	20.8
Soft Label	**46.5**	**31.2**	**31.7**	**24.8**

Cross-Camera Tracklet Association Learning. We evaluated the CCTA component by measuring the performance drop once eliminating it. Table 8 shows that CCTA brings a significant re-id accuracy benefit, e.g. a Rank-1 boost of 6.1% (49.9-43.8) and 12.1% (43.8-31.7) on MARS and DukeTracklet, respectively. This suggests the importance of cross-camera ID class correlation modelling and the capability of our CCTA formulation in reliably associating tracklets across cameras for unsupervised re-id model learning.

TABLE 8
Effect of Cross-Camera Tracklet Association (CCTA) learning.

Dataset	MARS [27]		DukeTracklet	
CCTA	Rank-1	mAP	Rank-1	mAP
✗	43.8	31.4	31.7	26.4
✓	**49.9**	**35.2**	**43.8**	**36.6**

To further examine why CCTA enables more discriminative re-id model learning, we tracked the self-discovered cross-camera tracklet matching pairs throughout the training. Figure 6 shows that both the number and precision of self-discovered cross-camera tracklet pairs increase. This echoes the model performance superiority of UTAL.

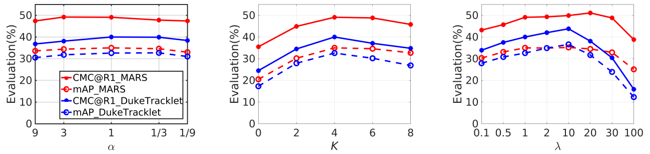


(a) Number of tracklet pairs. (b) Precision of matching pairs.

Fig. 6. The evolving process of self-discovered cross-camera tracklet matching pairs in **(a)** number and **(b)** precision throughout the training.

Model Parameters. We evaluated the performance impact of three UTAL hyper-parameters: (1) the tracklet feature update learning rate α (Eq (3)), (2) the sparsity K of the tracklet affinity matrix used in computing the soft labels (Eq (4) and (5)); (3) the loss balance weight λ (Eq (11)). Figure 7 shows that: (1) α is not sensitive with a wide satisfactory range. This suggests a stable model learning procedure. (2) K has an optimal value at "4". Too small values lose the opportunities of incorporating same-ID tracklets into the soft labels whilst the opposite instead introduces distracting/noisy neighbour information. (3) λ is found more domain dependent, with the preference values around "10".

This indicates a higher importance of cross-camera tracklet association and matching.



(a) The sensitive of α (b) The sensitive of K (c) The sensitive of λ

Fig. 7. Analysis of the UTAL model parameters.

Cross-Camera Nearest Neighbours (CCNN). We evaluated the effect of CCNN $\mathcal{R}$ (Eq (9)) used in the CCTA loss. We compared two designs: Our reciprocal 2-way 1-NN *vs.* common 1-way 1-NN. Table 9 shows that the more strict 2-way 1-NN gives better overall performance whilst the 1-way 1-NN has a slight advantage in Rank-1 on MARS (50.5% vs. 49.9%). By 2-way, we found using more neighbours (5/10-NN) degrades model performance. This is due to the introduction of more false cross-camera matches.

TABLE 9
Effect of cross-camera nearest neighbours.

Dataset	MARS [27]		DukeTracklet	
Metric (%)	Rank-1	mAP	Rank-1	mAP
1-way 1-NN	**50.5**	33.2	41.9	34.5
2-way 1-NN	49.9	**35.2**	**43.8**	**36.6**
2-way 5-NN	47.6	34.5	38.0	31.6
2-way 10-NN	45.5	32.5	33.9	27.1

Tracklet Sampling *versus* All Tracklets. In our preliminary solution TAUDL [28], we considered a Sparse Space-Time Tracklet (SSTT) sampling strategy instead of unsupervised learning on all tracklet data. It is useful in minimising the person ID duplication rate in tracklets. However, such a data sampling throws away a large number of tracklets with rich information of person appearance exhibited continuously and dynamically over space and time. To examine this, we compared the SSTT sampling with using all tracklets.

Table 10 shows two observations: (1) In overall re-id performance, our preliminary method TAUDL [28] is outperformed significantly by UTAL. For example, the Rank-1 results are improved by 6.1% (49.9-43.8) on MARS and by 17.7% (43.8-26.1) on DukeTracklet. (2) When using the same UTAL model, the SSTT strategy leads to inferior re-id rates as compared with using all tracklets. For instance, the Rank-1 performance drop is 4.8% (49.9-45.1) on MARS, and 12.1% (43.8-31.7) on DukeTracklet. These performance gains are due to the proposed soft label learning idea that effectively handles the trajectory fragmentation problem. Overall, this validates the efficacy of our model design in solving the SSTT's limitation whilst more effectively tackling the ID duplication (due to trajectory fragmentation) problem.

TABLE 10
Tracklet sampling *versus* using all tracklets.

Dataset	MARS [27]		DukeTracklet	
Metric (%)	Rank-1	mAP	Rank-1	mAP
TAUDL [28]	43.8	29.1	26.1	20.8
UTAL(SSTT)	45.1	31.1	31.7	24.8
UTAL(All Tracklets)	**49.9**	**35.2**	**43.8**	**36.6**

Effect of Neural Network Architecture. The model generalisation performance of UTAL may depend on the selection of neural network architecture. To assess this aspect, we evaluated one more UTAL variant using a more recent DenseNet-121 [78] as the backbone network, *versus* the default choice ResNet-50 [69]. Table 11 shows that even superior re-id performances can be obtained when using a stronger network architecture. This suggests that UTAL can readily benefit from the advancement of network designs.

TABLE 11
Effect of backbone neural network in UTAL.

Dataset	MARS [27]		DukeTracklet	
Metric (%)	Rank-1	mAP	Rank-1	mAP
ResNet-50 [69]	49.9	35.2	43.8	36.6
DenseNet-121 [78]	**51.6**	**35.9**	**44.3**	**36.7**

Weakly Supervised Tracklet Association Learning. For training data labelling in person re-id, the most costly procedure is on exhaustive manual search of cross-camera image/tracklet matching pairs. It is often unknown where and when a specific person will appear given complex camera spatio-temporal topology and unconstrained people's behaviours in the public spaces. Therefore, per-camera independent ID labelling is more affordable. Such labelled data are much weaker and less informative, due to the lack of cross-camera positive and negative ID pairs information. We call the setting *Weakly Supervised Learning* (WSL).

The proposed UTAL model can be flexibly applied in the WSL setting. Interestingly, this allows to test how much re-id performance benefit such labels can provide. Unlike in the unsupervised learning setting, the soft label based PCTD loss is *no longer* necessary in WSL given the within-camera ID information. Hence, we instead deployed the hard one-hot label (Eq (1)) based PCTD loss (Eq (2)). Table 12 shows that such weak labels are informative and useful for person re-id by the UTAL method. This test indicates a wide suitability and usability of our method in practical deployments under various labelling budgets.

TABLE 12
Evaluation of weakly supervised tracklet association learning.

Dataset	MARS [27]		DukeTracklet	
Metric (%)	Rank-1	mAP	Rank-1	mAP
Unsupervised	49.9	35.2	43.8	36.6
Weakly Supervised	**59.5**	**51.7**	**46.4**	**39.0**

Manual Tracking. *DukeMTMC-VideoReID* provides *manually* labelled trajectories, originally introduced for *one-shot* person re-id [66]. We tested UTAL on this dataset *without* the assumed one-shot labelled trajectory per ID. We set $K = 0$ for Eq (5) due to no trajectory fragmentation. Table 13 shows that UTAL outperforms EUG [66] even without one-shot ID labelling. This indicates the efficacy of our unsupervised learning strategy in discovering re-id information.

TABLE 13
Evaluation on DukeMTMC-VideoReID.

Metric (%)	Rank-1	Rank-5	Rank-20	mAP
EUG [66]	72.8	84.2	91.5	63.2
UTAL	**74.5**	**88.7**	**96.3**	**72.1**

5 Conclusions

We presented a novel *Unsupervised Tracklet Association Learning* (UTAL) model for unsupervised tracklet person re-identification. This model learns from person tracklet data automatically extracted from videos, eliminating the expensive and exhaustive manual ID labelling. This enables UTAL to be more scalable to real-world applications. In contrast to existing re-id methods that require exhaustively pairwise labelled training data for every camera-pair or assume labelled source domain training data, the proposed UTAL model performs end-to-end deep learning of a person re-id model from scratch using totally unlabelled tracklet data. This is achieved by optimising jointly both a Per-Camera Tracklet Discrimination loss function and a Cross-Camera Tracklet Association loss function in a unified architecture. Extensive evaluations were conducted on eight image and video person re-id benchmarks to validate the advantages of the proposed UTAL model over state-of-the-art unsupervised and domain adaptation re-id methods.

Acknowledgments

This work is partially supported by Vision Semantics Limited, National Natural Science Foundation of China (Project No. 61401212), Royal Society Newton Advanced Fellowship Programme (NA150459), and Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149).

References

[1] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person re-identification*. Springer, 2014.

[2] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. Int. Jo. Conf. of Artif. Intell.*, 2017.

[3] ——, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2285–2294.

[4] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 79–88.

[5] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1179–1188.

[6] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2109–2118.

[7] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang, "Deep group-shuffling random walk for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2265–2274.

[8] H. Wang, S. Gong, and T. Xiang, "Unsupervised learning of generative topic saliency for person re-identification," in *Proc. Bri. Mach. Vis. Conf.*, 2014.

[9] E. Kodirov, T. Xiang, and S. Gong, "Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification," in *Proc. Bri. Mach. Vis. Conf.*, 2015.

[10] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1629–1642, 2015.

[11] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Person re-identification by unsupervised l_1 graph learning," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 178–195.

[12] F. M. Khan and F. Bremond, "Unsupervised data association for metric learning in the context of multi-shot person re-identification," in *Proc. IEEE Conf. Adv. Vid. Sig. Surv.*, 2016, pp. 256–262.

[13] H. Wang, X. Zhu, T. Xiang, and S. Gong, "Towards unsupervised open-set person re-identification," in *IEEE Int. Conf. on Img. Proc.*, 2016, pp. 769–773.

[14] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong, "Person re-identification by unsupervised video matching," *Pattern Recognition*, vol. 65, pp. 197–210, 2017.

[15] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen, "Dynamic label graph matching for unsupervised video re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5142–5150.

[16] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2429–2438.

[17] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2275–2284.

[18] P. Peng, Y. Tian, T. Xiang, Y. Wang, M. Pontil, and T. Huang, "Joint semantic and latent attribute modelling for cross-class transfer learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1625–1638, 2018.

[19] H. Fan, L. Zheng, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *arXiv:1705.10444*, 2017.

[20] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 994–1002.

[21] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 152–159.

[22] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1116–1124.

[23] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Workshop of Eur. Conf. Comput. Vis.*, 2016, pp. 17–35.

[24] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3754–3762.

[25] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 688–703.

[26] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scand. Conf. Img. Anal.*, 2011, pp. 91–102.

[27] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 868–884.

[28] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 737–753.

[29] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 392–408, 2018.

[30] H. Wang, S. Gong, X. Zhu, and T. Xiang, "Human-in-the-loop person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 405–422.

[31] C. Liu, C. Change Loy, S. Gong, and G. Wang, "Pop: Person re-identification post-rank optimisation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 441–448.

[32] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2360–2367.

[33] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3586–3593.

[34] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3550–3557.

[35] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[36] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 994–1003.

[37] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero-and homogeneously," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 172–188.

[38] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2903058, IEEE Transactions on Pattern Analysis and Machine Intelligence

re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1306–1315.

[39] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7948–7956.

[40] S. Bak, P. Carr, and J.-F. Lalonde, "Domain adaptation through synthesis for unsupervised person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 189–205.

[41] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. Neur. Info. Proc. Sys.*, 2007, pp. 41–48.

[42] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[43] Q. Dong, S. Gong, and X. Zhu, "Multi-task curriculum transfer deep learning of clothing attributes," in *Proc. IEEE Win. Conf. App. of Comp. Vis.*, 2017, pp. 520–529.

[44] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 918–930, 2016.

[45] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[46] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 478–487.

[47] P. Bojanowski and A. Joulin, "Unsupervised learning by predicting noise," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 517–526.

[48] Z. Wu, Y. Xiong, X. Y. Stella, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.

[49] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2794–2802.

[50] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[51] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1259–1267.

[52] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "Motchallenge 2015: Towards a benchmark for multi-target tracking," *arXiv:1504.01942*, 2015.

[53] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journ. of Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.

[54] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. Neur. Info. Proc. Sys.*, 2005, pp. 1601–1608.

[55] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[56] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *Proc. Int. Conf. on Learn. Rep.*, 2017.

[57] D. Qin, S. Gammeter, L. Bossard, T. Quack, and L. Van Gool, "Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 777–784.

[58] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[59] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[60] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[61] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area v2," in *Proc. Neur. Info. Proc. Sys.*, 2008, pp. 873–880.

[62] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1249–1258.

[63] J. Ye, Z. Zhao, and H. Liu, "Adaptive distance metric learning for clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–7.

[64] C. Qin, S. Song, G. Huang, and L. Zhu, "Unsupervised neighborhood component analysis for clustering," *Neurocomputing*, vol. 168, pp. 609–617, 2015.

[65] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep crf for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8649–8658.

[66] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5177–5186.

[67] M. Gou, S. Karanam, W. Liu, O. Camps, and R. J. Radke, "Dukemtmc4reid: A large-scale multi-camera person re-identification dataset," in *Workshop of IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 10–19.

[68] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[69] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[71] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 170–186.

[72] Y. Chen, X. Zhu, and S. Gong, "Deep association learning for unsupervised video person re-identification," *Proc. Bri. Mach. Vis. Conf.*, 2018.

[73] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1169–1178.

[74] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asi. Conf. Comp. Vis.*, 2012, pp. 31–44.

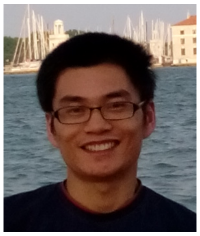
[75] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 262–275.

[76] B. J. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. Bri. Mach. Vis. Conf.*, 2010.

[77] D. J. Pearce, "An improved algorithm for finding the strongly connected components of a directed graph," *Victoria University, Wellington, NZ, Tech. Rep*, 2005.

[78] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

Minxian Li is a postdoctoral researcher at Queen Mary University of London, United Kindom and also an assistant professor of Nanjing University of Science and Technology, China. He received his Ph.D. in Pattern Recognition and Intelligent System from Nanjing University of Science and Technology, China. His research interests include computer vision, pattern recognition and deep learning.



Xiatian Zhu is a Computer Vision Researcher at Vision Semantics Limited, London, UK. He received his Ph.D. from Queen Mary University of London. He won The Sullivan Doctoral Thesis Prize 2016, an annual award representing the best doctoral thesis submitted to a UK University in computer vision. His research interests include computer vision and machine learning.

Shaogang Gong is Professor of Visual Computation at Queen Mary University of London (since 2001), a Fellow of the Institution of Electrical Engineers and a Fellow of the British Computer Society. He received his D.Phil (1989) in computer vision from Keble College, Oxford University. His research interests include computer vision, machine learning and video analysis.